DOCUMENT RESUME

ED 420 693 TM 028 368

AUTHOR Johanson, George; Green, Suzy; Williams, James
TITLE Blunders and Missed Opportunities in Survey Research.
PUB DATE 1998-04-00
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Questionnaires; Research Design; *Research Methodology; Research Problems; *Surveys
IDENTIFIERS Outliers

ABSTRACT

Errors in methodology occur regularly in the conduct of surveys for educational research. This paper discusses some of these errors and alternatives. In the area of survey design, errors are common in: (1) missed opportunities in attitude scale planning; (2) blunders in item construction; (3) missed opportunities in item construction; (4) blunders in scale revision; and (5) missed opportunities in scale revision. In the area of survey analysis, there are blunders and missed opportunities in outlier disposition, as there are in nonresponse. With regard to the interpretation of survey results, blunders in causal conclusions and missed opportunities with true experiments are common. This list is far from complete, but it does expose some of the more blatant errors in survey research. A researcher cannot adequately correct earlier errors with later procedures, and so should be attentive throughout the entire survey research process. (Contains 17 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Blunders and Missed Opportunities in Survey Research

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

George Johanson, Suzy Green, and James Williams
College of Education
Ohio University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
George Johanson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

A Paper Presented at the 1998 Annual Meeting of the
American Educational Research Association, San Diego, CA.

ED 420 693

TM028368

Our experience in conducting, consuming, and guiding educational research using survey methodology confirms our current contention that there are errors of commission and omission occurring regularly. The literature on survey research methods addresses many of the errors we identify, but these particular errors may not be as widely known as desired and seem to be repeatedly subject to misunderstanding. The primary purpose of our paper is to reveal at least some of these errors and note alternatives.

We have structured our concerns and comments around blunders and missed opportunities within three main stages of survey research: survey design, analysis, and interpretation.

Survey Design

Missed Opportunities in Attitude Scale Planning

A missed opportunity in planning occurs when the researcher fails to use multiple sources of information to determine the components of an attitude of interest. The use of interviews with a small convenience sample of respondents and/or experts-in-the-field to identify the various components of an attitude is often recommended. Converse and Presser (1986) recommend that "...interviewing even a few individuals can enrich the researcher's perspective." (p. 49).

Others prefer to use the literature for identifying the components of the attitudinal construct under investigation. Spector (1992) notes that "The existing literature should serve as a starting point for construct definition." (p. 14). However, it is only when the attitudinal components of the construct located in the literature confirm those components from interviews with experts and/or a convenience sample of subjects, that there is assurance that unnecessary components have been avoided and that critical aspects of the attitude have not been overlooked. It seems to us that this form of triangulation is often overlooked. It is the *convergence* of the literature and information from experts (or a small sample of subjects) that gives the researcher the confidence that the attitudinal construct has been accurately identified and the opportunity for a complete and precise definition of the attitude in question has not been missed.

Blunders in Item Construction

After the pool of items for an attitude scale has been initially drafted, a small group of subjects can be recruited to provide additional information about these items. Sudman, Bradburn, and Schwarz (1996) declare: "In questionnaire design, we strongly recommend the use of thinkaloud interviews for determining what respondents think the questions mean and how they retrieve information to form a judgment." (p. 53). A major purpose of having a small sample of subjects 'think aloud' while reading and reflecting on the items is to better understand the cognitive processes involved in making the required responses to the items. At a very practical level, the researcher using such methods for attitude scale development should be able to

determine if the survey items are likely to perform as intended or if a blunder is in the offing.

As a simple example, consider a survey question such as: *How many miles do you commute to work each day?* If a small number of respondents are asked to think aloud as they respond to this item, it will quickly become clear that some are interpreting this to mean *one-way* mileage while others are giving *round-trip* estimates. This critical observation might be overlooked with only a brief written pilot study. Of course, survey items that require multiple steps or even some problem solving might best use these methods. Someren et al., 1994, is a good source of information for think aloud methods.

A blunder of a different sort can occur when using Likert scaling. When reviewing a set of Likert items, it is critical to note whether each item is positively or negatively phrased. Our experience has been that less experienced researchers become so involved in the creative portion of item-writing that often a clearly neutral item will creep into the longer scales. In, say, an *Attitude Towards Chocolate* scale, it is a blunder to ask respondents if they strongly agree that 'Chocolate tastes OK'.

Missed Opportunities in Item Construction

Should a researcher construct items that force choices? While there are instances where a forced-choice format would be less appropriate, there are also situations where this is definitely not true. For instance, if you request faculty and staff in a public school to indicate preferences for the next year's school calendar, you had better force some choices or you will end up with data indicating that the sentiment was strongly in favor of starting school later, taking longer vacations, and having an earlier closing date! Unfortunately, this missed opportunity was not apocryphal, but a first-person (mercifully unpublished) experience of one of the authors many years ago.

Offering a neutral, unsure, or don't know (DK) option can also make a good deal of sense, but not in all circumstances. The middle option may sometimes serve as an opportunity for respondents to avoid responding to questions they would rather not answer. On personnel evaluations, the number of DKs can sometimes be observed to correlate significantly (and negatively) with the overall rating of the individual be evaluated (Johanson, Gips, & Rich, 1993).

Still another item format, the free-response or open-ended format, can be profitably used either at the beginning or end of a survey with many closed or selected-response items. However, the item may function very differently depending on its position. If used prior to the more structured items, the open-ended item will permit the researcher to determine whether there are concepts or ideas that were missed or omitted in the structured response format items. Such an open-ended item can provide evidence of content validity, when phrased something like 'What is the most important aspect of your attitude towards...' and if the responses can be classified or clustered to reflect the same ideas or constructs as the closed-response items. If used at the end of a survey, the open-ended item cannot serve this purpose since the respondent will most certainly have been influenced by the prior closed-response items. Nonetheless, free-response items at the end of the survey phrased as 'Do you feel there are any additional concepts or concerns...' or 'Were there items that seemed less appropriate...' may also identify a missed concept or opportunity, especially in the earlier stages of scale development.

The use of an overall or summary item in a survey is often discouraged. An overall item in an attitude scale should be avoided because such items require the respondent to define or specify the content or meaning of the item and this meaning will differ from person to person. In this instance, the researcher would know the response but not the actual question that prompted the

response. This would not be very helpful. Authors such as Schuman & Presser (1996) feel that: "General summary type items are especially susceptible to context effects and should probably be avoided if the needed information can be built up from more specific questions." (p. 311). It would seem there are at least several good reasons for not using summary or overall items.

Nevertheless, if such an item is the last item in an attitude scale with more specific items, the overall item can serve at least two very useful purposes. First, an item analysis will often identify the summary item as the strongest item (in the sense of the highest item-total correlation) in the scale! Second, the responses to this overall item can be used as a surrogate for the overall score and, for example, be correlated with the responses to a suspect item to see if an item of suspect validity is at least positively related to the central idea of the scale. This may be especially useful in the developmental stages of a scale where the summated score itself may reflect many flawed items.

Last in this collection of missed opportunities is the often used ranking task. Ranking can require much effort on the part of respondents to do correctly and may often be done with too little thought. How is a researcher to know whether the subjects have given sufficient time and attention to the necessary comparisons in a typical ranking chore? One alternative is to present the objects to be ranked as all possible pairs and require the respondent to make separate pairwise choices. While there will be a large number of these individual pairwise items (there will be k(k-1)/2 pairs where k is the number of objects to be ranked), there are advantages to this approach.

The first advantage is that we have structured the task to insure that all comparisons have been considered; the responses to the paired-comparisons will still yield an overall ranking. Second, the responses can be scaled in a variety of ways including some that yield approximately interval level measures such as Thurstone scaling (Dunn-Rankin, 1983). Finally, the number of logical inconsistencies or *circular triads* or *intransitivities* can be computed. These inconsistencies might occur where a respondent indicates a preference for A over B and B over C, but then, inexplicably, chooses C over A. Information on the number of circular triads may reflect the scalability of the objects or the attention of the respondents to the task. Formulas for computing the number of circular triads for each subject are in Kendall (1970). Using paired-comparisons instead of ranking requires more items, but provides several opportunities for added information.

Blunders in Scale Revision

Pilot studies are conducted, in part, to identify poorly functioning items and permit attitude scale revisions prior to actual data collection. What is a researcher to do if the efforts at scale development were somehow flawed or too abbreviated and expensive data have subsequently been assembled using an instrument with some obviously imperfect items? It is very tempting to identify the flawed items and then reanalyze the results with a newly revised or shortened scale. This is the classic *foldback*. It is very seductive, can lead to some very dubious results, and should be avoided. It is rather easy to demonstrate that selecting items that intercorrelate reasonably well *from randomly generated data* will yield an acceptably reliable scale when the reliability analysis is folded back on the original data from which the items were selected. For instructional purposes, student responses to nonsense items can also be used to illustrate this point.

The enticement for the researcher where data have already been collected is not limited to reliability analyses. When a factor analysis identifies items which seem to load inappropriately, or not at all, it is too easy to drop them and proceed as usual. To be succinct, *all significant scale*

revision requires cross-validation. If instrument changes absolutely have to be made with a final data set, then the researcher should cross-validate by randomly splitting the data set, using one half to identify the revisions, and computing results on the other half.

A second blunder in scale revision is related to internal consistency reliability. There is a temptation to claim that a high internal consistency implies a unidimensional scale. For example, Mueller (1986) states: "If a scale has a high index of internal consistency, we know that the items are substantially intercorrelated. They are working together to measure the same underlying variable. This constitutes evidence that a construct is being measured." (p. 71). Unfortunately, this is easily challenged by constructing two reasonably uncorrelated unidimensional scales and noting that the internal consistency reliability of the combined (two-dimensional) scale is often quite acceptable. Simply put, evidence of reliability, by itself, does not provide evidence of either unidimensionality or construct validity. This blunder can also be a missed opportunity if a reliability coefficient discourages further efforts to investigate dimensional structure.

Missed Opportunities in Scale Revision

In addition to a usual item analysis, it may be wise in some instances to check as well for *differential item functioning*, or DIF. That is, if there are two groups of respondents it is possible for one item to be more easily agreed-with by members of one group even after controlling for overall attitude level. Items that function differentially can lead to biased measurement. A more complete discussion of DIF is available in Camilli & Shepard, 1994. Johanson (1997) discusses applications of DIF techniques to attitudinal data and gives an example using a science attitude scale in which a dinosaur item was functioning differentially for boys and girls and would have been biased for many applications.

Analysis

Blunders in Outlier Disposition

Exceptional data values cannot be avoided and it is certainly shortsighted to automatically ignore or drop such cases without even a modest amount of exploration. However, in some situations, a response or group of responses may be unreasonable. A response to a closed format item that is different from all possible responses is either unreasonable data or uncertain data entry. If a subject on a survey indicates an exceptional number of years of experience in a profession, then the respondent's age should be consistent with this experience. In fact, we have seen instances where such a difference (age minus years of professional experience) was negative! In both situations, checks on data accuracy were needed.

Missed Opportunities in Outlier Disposition

Cases presenting exceptional or seemingly contradictory values may be some of the most interesting and informative in any data set. Carefully studying the responses of such subjects is much more reasonable than omitting them. Who are the subjects with these values? What comments did they make? What characteristics do they share? Are the results of the data analyses similar both with and without these exceptional subjects? Perhaps the best way to 'see' the impact of an outlier or an influential data point is to plot the relevant variable and relationship.

Omitting responses or even dropping a respondent is much less sensible in situations where responses are extreme but still within the realm of possibility. With a reasonably large sample size, a few extreme responses are expected. It is only when the results of data analyses are similar with outliers both removed and included that the decision on outlier disposition is less than critical. Discarding unusual, but possible, data points can lead to missed opportunities, but

retaining impossible data points is a blunder.

Blunders in Nonresponse

We often spend considerable time and attention in our statistics instruction on the importance of having sufficient statistical power to detect desired effects with reasonable probability. Our students seem to understand the need for an adequate number of respondents, but it is frustrating and costly to pursue nonrespondents. There is a temptation to bolster sample size by returning to the population for a larger sample when the response rate is less than anticipated rather than relentlessly pursuing nonrespondents from the original sample. Unfortunately, larger samples of only the most compliant subjects may be substantially less representative of a population than a smaller randomly selected sample. Following up on nonrespondents is difficult and costly, but there is not a blunder-free alternative.

Missed Opportunities in Nonresponse

Missing data in mailed surveys caused by unit nonresponse is a difficult problem. Recommendations are often based on avoiding missing data and not on dealing with missing data. Still, even the shortest survey with the best follow-ups and procedures will likely result in some unit nonresponse. Some authors suggest sampling the nonrespondents (Tuckman, 1994), but this is difficult since these are persons who have previously resisted all attempts to elicit a response.

Our suggestion is also to sample the nonrespondents, but only with the single most salient item in your survey and do it with a postcard response. In a study of secondary school principals (Johanson & Gips, 1992), over 50% of the nonrespondents to two previous mailings were willing to respond to such an appeal and provided a response to a single item. This rate of response from the previously nonresponding subjects was very similar to the overall response rate of the entire group and yielded an overall reply of approximately 75% to the critical item. Perhaps more importantly in this study, these more reluctant respondents gave a somewhat different mean rating to this single item which limited the generalizability of the findings.

Interpretation

Blunders in Causal Conclusions

The convenient error of drawing causal conclusions in research studies with only correlational data is sufficiently well known to have a name: the *post hoc* fallacy (Best & Kahn, 1993). This mistake is certainly not unique to those using survey research methods.

A less blatant sort of *post hoc* fallacy occurs when a researcher acknowledges the correlational nature of the data (perhaps even attesting to the lack of a valid causal argument), but then makes recommendations for practice that are sensible only if there is a causal link between the variables under investigation. A common instance of this would be a correlational study that involves both an attitude and an outcome measure; the recommendation for practice is the implementation of programs for the purpose of attitude adjustment. Such suggestions for practice make no sense without the underlying blunder of a causal assumption.

Missed Opportunities with True Experiments

Many educational researchers would likely agree with Wiersma (1995) that "Survey research is undoubtedly the most widely used nonexperimental type of educational research." (p. 206). Interestingly, it is probably easier to conduct true experiments (using random assignment of subjects to treatments) with mailed surveys than with many other forms of research. The problem is that many of us seem to feel that such experimental research is limited to survey research methodology on such topics as phrasing differences or context effects.

Note, however, that the use of *vignettes* (Converse & Presser, 1986) or brief 'story' items makes this experimental approach to survey research both simple and versatile. For example, if a researcher wishes to determine the nature and extent of hiring biases, it is rather straight forward to construct an abbreviated résumé where only the variable of interest (e. g., gender) is different across two or more forms of the survey. If respondents are randomly assigned to treatments (randomly sent only one of the survey forms), then systematic mean résumé evaluation differences between the forms must be due to the only manipulated variable (in this case, gender). Opportunities for experimental survey research would seem to be missed more often than necessary when survey research is portrayed as exclusively correlational.

Discussion

Our list of blunders and missed opportunities is far from complete (issues relating to survey sampling and nonparametric statistics, for example, were not even mentioned). We have attempted only to expose certain of the more blatant errors in survey research that we routinely encounter. A critical observation in considering methodological errors of any sort is the interdependence of the steps or stages in survey research. A blunder or missed opportunity at the design phase may well result in additional, or compounding, errors at the analysis or interpretation stages of the research.

Does this imply that the earlier errors are more insidious or corrupting? We do not think that this is the fundamental problem or the message to be delivered. It is true that errors in the construction phase of an attitude scale will ripple throughout the research process and that decisions made about extreme data points may well influence later interpretations. Yet, neither mistake may be as misleading to the consumer as a subtle *post hoc* fallacy in the conclusion.

Our rather simple point is that a researcher cannot adequately correct earlier errors with later procedures and so needs to be knowledgeable and attentive throughout the entire survey research process. Many of these blunders and missed opportunities can be happily avoided merely by being forewarned and anticipating problem areas.

References

Best, J. W., & Kahn, J. V. (1993). Research in education (7th ed.). Needham Heights, MA: Allyn and Bacon.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.

Converse, J. M., & Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. Beverly Hills, CA: Sage Publications.

Dunn-Rankin, P. (1983). Scaling methods. Hillsdale, NJ: Lawrence Erlbaum.

Johanson, G. A. (1997). Differential item functioning in attitude assessment. Evaluation Practice, 18(2), 127-135.

Johanson, G. A., Gips, C. J., & Rich, C. E. (1993). "If you can't say something nice": A variation on the social desirability response set. Evaluation Review, 17(1), 116-122.

Johanson, G. A. & Gips, C. J. (1992). Secondary school hiring practices. The High School Journal, 76(1), 1-16.

Kendall, M. B. (1970). Rank correlation methods (4th ed.). New York: Hafner Press.

Mueller, D. J. (1986). Measuring social attitudes. New York: Teachers College, Columbia University.

Schuman, H., & Presser, S. (1996). Questions & answers in attitude surveys: Experiments on question form, wording, and context. Thousand Oaks, CA: Sage Publications.

Schwarz, N., & Sudman, S. (Eds.) (1996). Answering questions: Methodology for determining cognitive and communicative processes in survey research. San Francisco, CA: Jossey-Bass.

Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). The think aloud method: A practical guide to modelling cognitive processes. London: Academic Press.

Spector, P. E. (1992). Summated rating scale construction: An introduction. Beverly Hills, CA: Sage Publications.

Sudman, S., & Bradburn, N. (1982). Asking questions: A practical guide to questionnaire design. San Francisco, CA: Jossey-Bass.

Sudman, S., Bradburn, N. , & Schwarz (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco, CA: Jossey-Bass.

Tuckman, B. W. (1994). Conducting educational research (4th ed.). Fort Worth, TX: Harcourt Brace.

Wiersma, W. (1995). Research methods in education: An introduction (6th ed.). Needham Heights, MA: Allyn and Bacon.